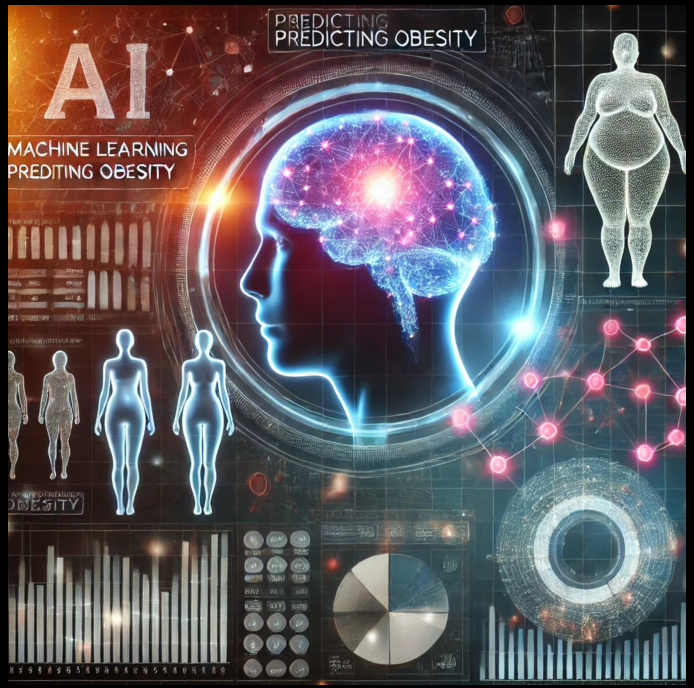


Predicting Obesity with Machine Learning

Elham Rahimi

Introduction

- Obesity is a global health concern with many medical implications
- Early predictions can help prevent obesity
- Machine learning lets us use data to predict obesity risk based on lifestyle and eating habits.

Project Objectives

- To develop a machine learning model to predict the likelihood of a person becoming obese
- Identify factors that contribute to the likelihood of a person becoming obese
- With this model; meal planners, diet centers and gyms can use this model to help their customers.

Usage

Meal Planning Services

Diet Centers

Gym

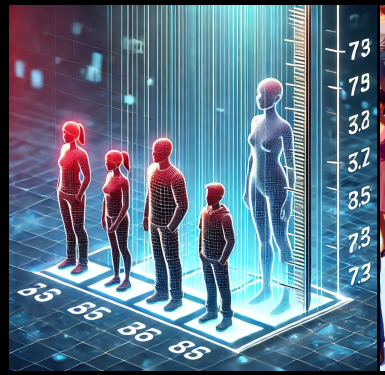
Doctors

Dataset

- Kochar, J. P. (n.d.). *Obesity Risk Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/jpkochar/obesity-risk-dataset

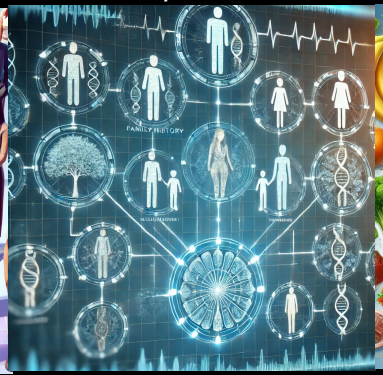
The features in the dataset include:

Anthropometric Features



Demographic Features

Genetic & Family History Features

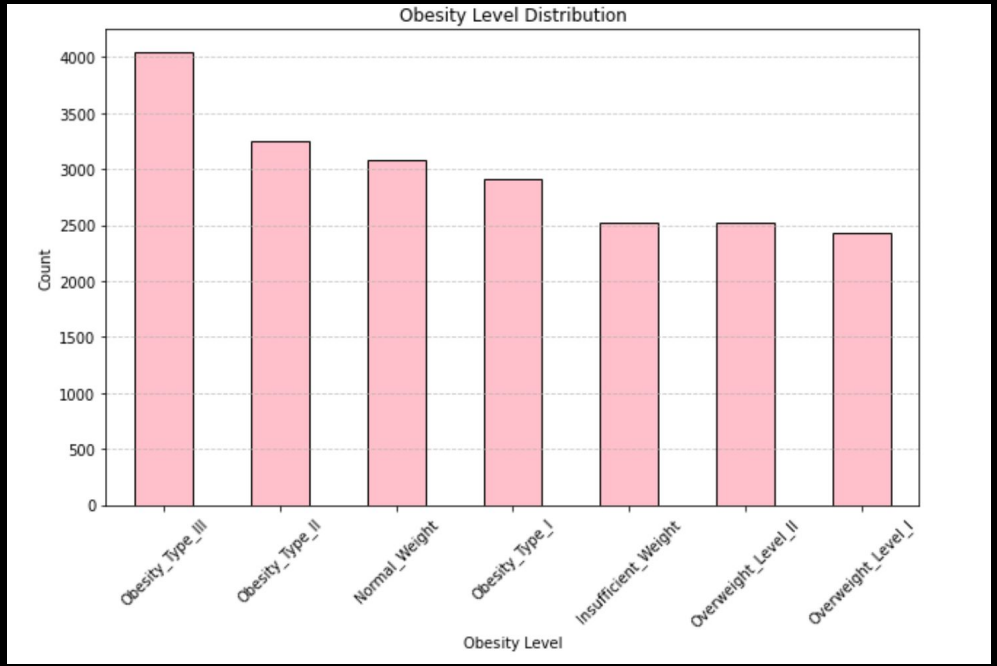


Dietary Habits & Nutrition Features

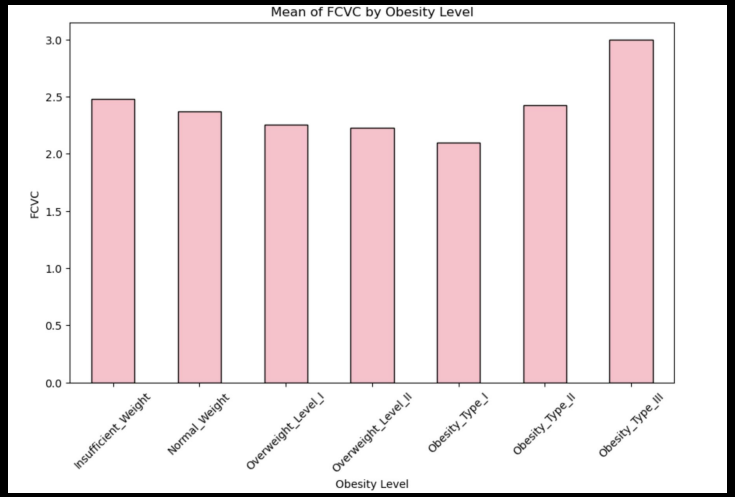
Lifestyle Features

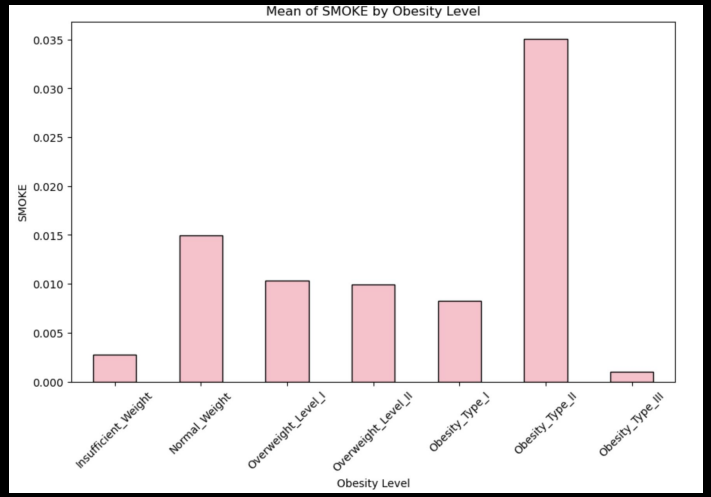
Distribution of Obesity Level by BMI

Obesity Level	Estimated BMI
Insufficient Weight	〈 18.5
Normal Weight	18.5 – 24.9
Overweight Level I	25 – 29.9
Overweight Level II	30 – 34.9
Obesity Type I	35 – 39.9
Obesity Type II	40 – 44.9
Obesity Type III	〉45



Findings from EDA

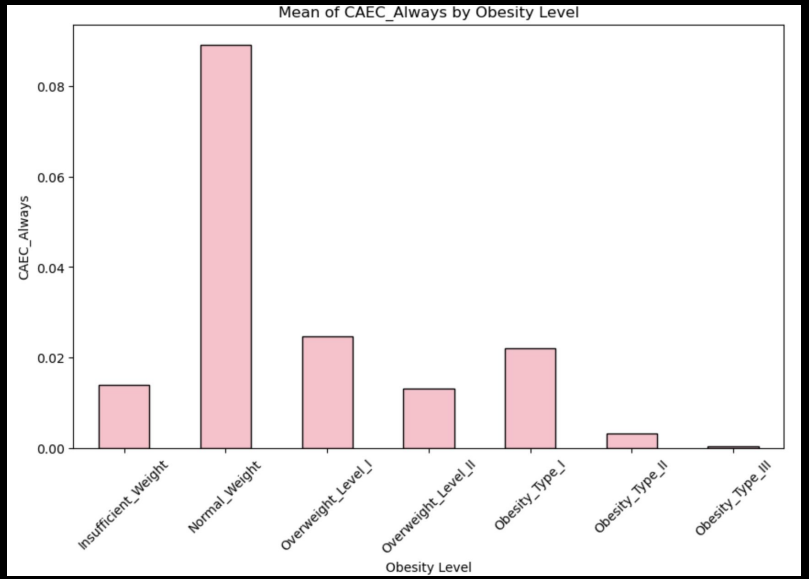




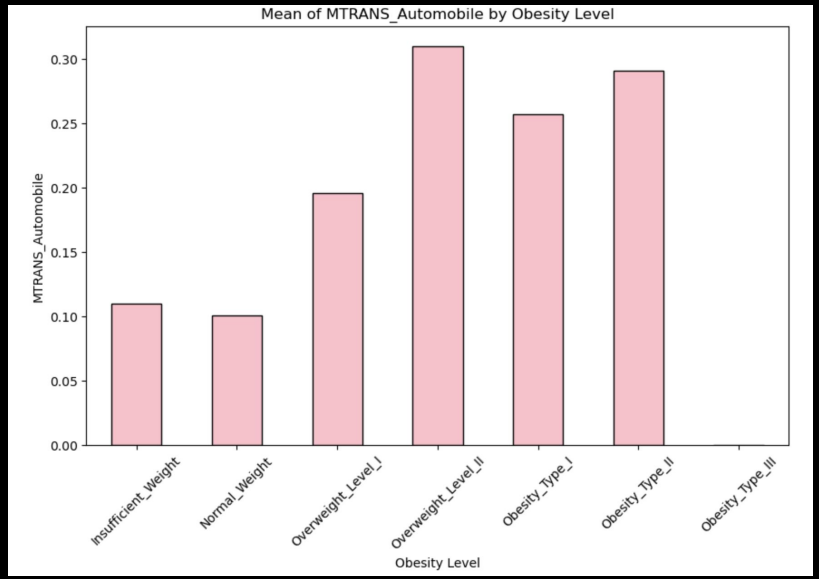
People with high obesity levels consumed more vegetables than low levels

Type II obesity had the highest smokers, could indicate that food could be addictive like smoking

Findings from EDA



Snacking between meals was most common in normal levels

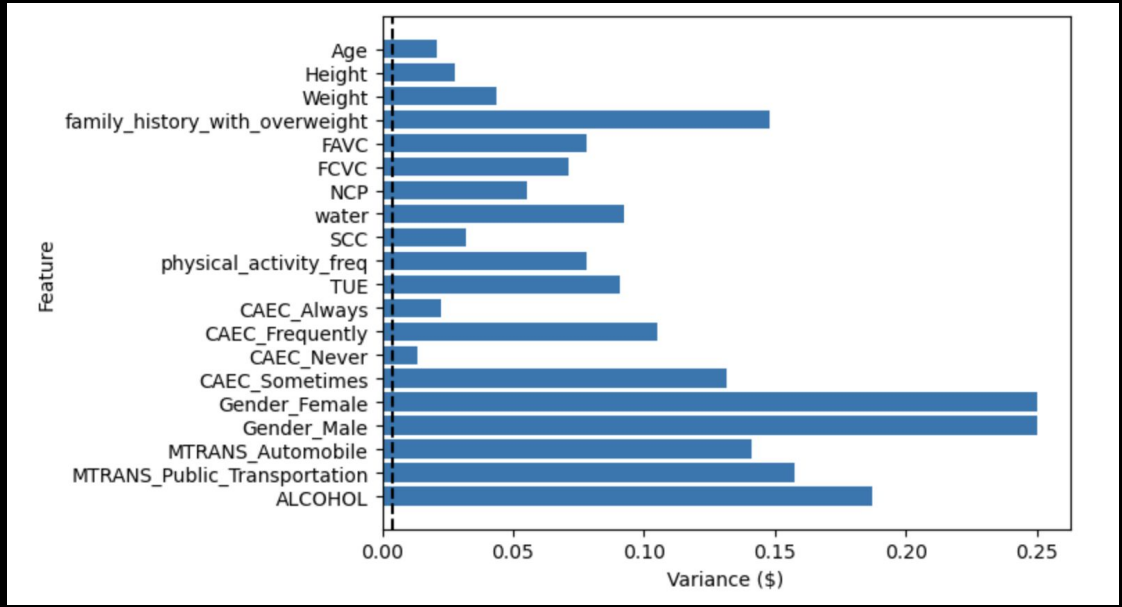


High obesity levels travelled mostly by automobile, whereas the low levels travelled by bicycles or motorcycles

Feature Engineering

The following was done to clean the dataset

1. Combines rare one-hot columns to reduce sparsity
2. Created a binary alcohol feature (similar to smoke)
3. Dropped smoke and some transportation features that barely varied across the dataset
4. Checked for low variance features to remove (threshold - 0.001)



Grouped PCA

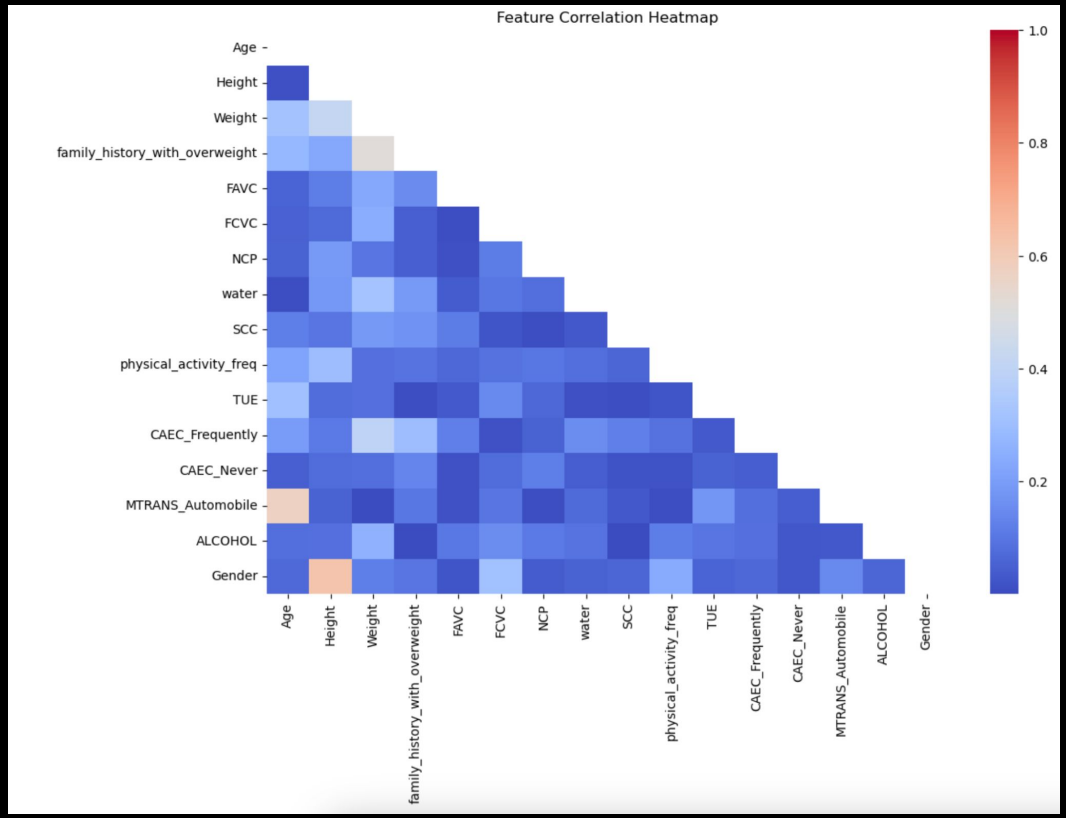
Grouped PCA is when you split your features into categories and run PCA on each group separately instead of on the whole dataset. It helps reduce dimensionality while still keeping some structure.

Group #	Group Name	Features
1	Demographics	Age, Height, Weight, Gender_Female, Gender_Male
2	Family History	family_history_with_overweight
3	Diet / Nutrition	FAVC, FCVC, NCP, CAEC_Always, CAEC_Frequently, CAEC_Never, CAEC_Sometimes, SCC
4	Alcohol Consumption	Alcohol
5	Water Consumption	water
6	Physical Activity	physical_activity_freq
7	Transportation & Tech	MTRANS_Automobile, MTRANS_Public_Transportation, TUE

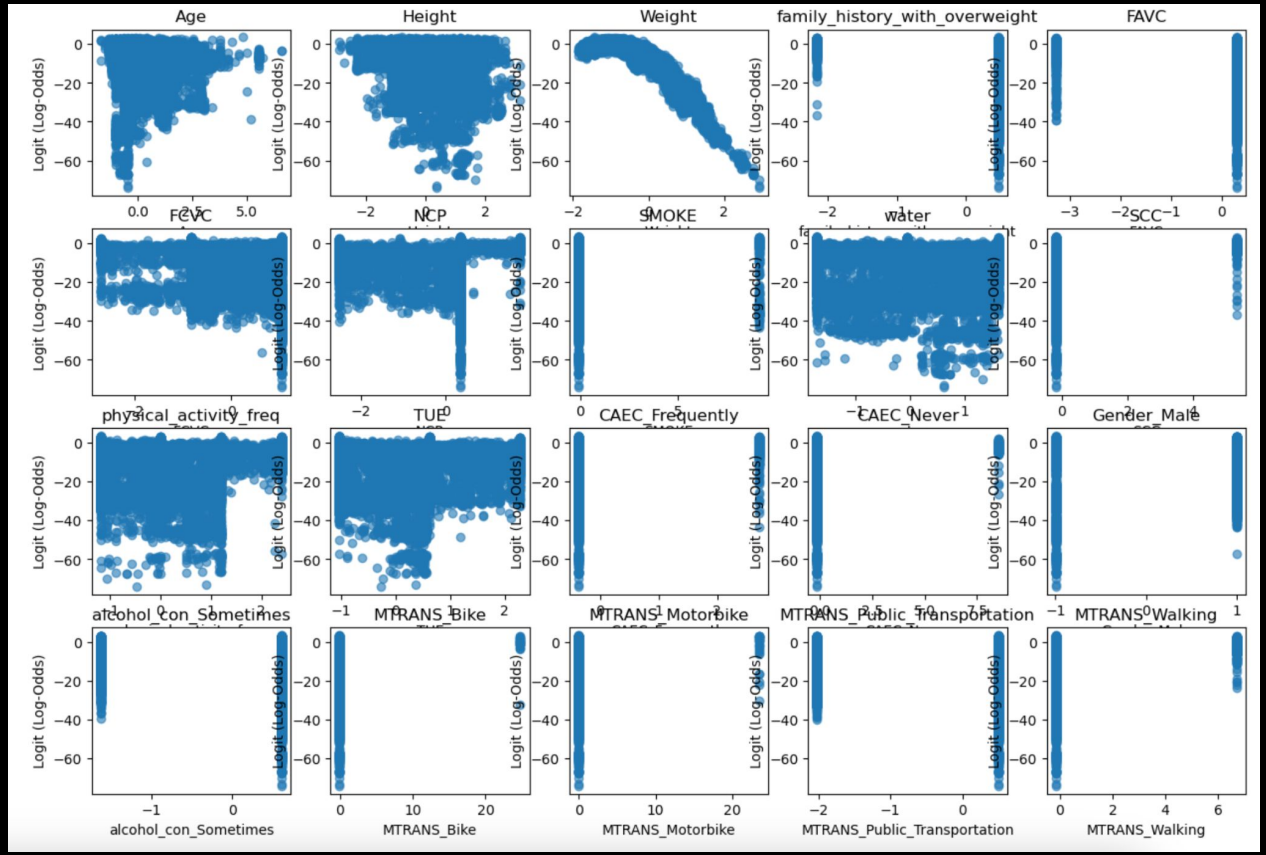
Baseline Model
Model 1: Logistic Regression

Preventing overfitting by checking for multicollinearity

* After dropping 'CAEC_Always','CAEC_Sometimes','MTRANS_Public_Transportation'



Making sure there's linearity between features and log of odds



Logistic Regression

- Model performed best without grouped PCA and PCA
- With pipeline and hyperparameter tuning the best parameters were:
  - C = 10
  - Penalty = ‘l1’
  - Solver = ‘saga’
- Accuracy : 87.19%

Logistic Regression: Model Evaluation

	precision	recall	f1-score	support
Insufficient_Weight	0.89	0.94	0.91	524
Normal_Weight	0.86	0.82	0.84	626
Obesity_Type_I	0.83	0.84	0.84	543
Obesity_Type_II	0.96	0.97	0.97	657
Obesity_Type_III	1.00	1.00	1.00	804
Overweight_Level_I	0.73	0.74	0.73	484
Overweight_Level_II	0.73	0.69	0.71	514
accuracy			0.87	4152
macro avg	0.86	0.86	0.86	4152
weighted avg	0.87	0.87	0.87	4152

- The model achieved 87% accuracy, with good precision and recall across most obesity classes.
- It performed best on Obesity_Type_III and II, while Overweight levels were harder to separate.
- Overall, the scores are well-balanced, with a macro F1 of 86%.

Baseline Models
Model 2: Decision Trees

Decision Trees:

- With pipeline and hyperparameter tuning the best parameters were:
  - Criterion = ‘entropy’
  - Max Depth = 9
  - Min_samples_leaf = 20
  - Min_samples_split = 12
- Accuracy : 87.60%

	Logistic Regression	Decision Trees
Pros	- Easy, fast and simple	- Handles nonlinearity - Can tell the most important features - Doesn’t require Scaling
Cons	- Handles linearity - Doesn’t tell the important features	- Not smooth boundary lines
Accuracy	87.19 %	87.60 %

Next Steps:

Moving forward, I'd like to explore Random Forest as a next model. Since it's an ensemble method that builds on decision trees, it could help improve accuracy and reduce overfitting by averaging the results of multiple trees. It also handles feature importance well, which would give me even more insight into which lifestyle factors are driving obesity predictions. I'm curious to see how it performs compared to logistic regression.